85. Deutscher Archivtag 2015, Karlsruhe

Fortbildungsveranstaltung 5: Computerwerkzeuge im Archiv von Bewertung bis Benutzung

Infoblatt „Webarchivierung“

Webarchivierung wird in den meisten Fällen unter Verwendung von Webcrawlern durchgeführt. Das am weitesten verbreitete Programm ist **Heritrix.** Er wird mittels einer Crawler-Beans.cxml konfiguriert. Die Konfiguration kann im **Heritrix User Guide** nachgelesen werden. Wichtig bei der Durchführung eines Crawls sind die Angabe eines korrekten Seeds und die Beachtung technischer Restriktionen.

Die zu archivierende Website wird von Heritrix im **warc**-Format gespeichert. Dies ist der de facto-Standard im Bereich der Webarchivierung.

Die so entstandenen warc-Dateien werden zum Beispiel mit **pyWB** wiedergegeben. Eine Alternative zu pyWB ist die **Wayback Machine**, die vom Internet Archive entwickelt wurde. Für die Wiedergabe werden die in der archivierten Website enthaltenen Hyperlinks umgeschrieben, sodass sie auf die archivierte Ressource verweisen.

Zur Vorabanalyse der zu archivierenden Website kann das Online-Tool **ArchiveReady** verwendet werden. Das Tool prüft eine Ressource unter einer gegebenen URL auf ihre Archivierungsfähigkeit und listet zu erwartende Probleme bei der Spiegelung und Archivierung auf.

Probleme ergeben sich unter anderem, wenn die betreffende Website Datenbanken oder Inhalte, die außerhalb der Domain liegen (z.B. GoogleMaps oder YouTube), verwendet. Auch ein umfänglicher Einsatz von Javascript oder Flash kann das Spiegelungsergebnis negativ beeinflussen.

Alle im Workshop verwendeten Materialien finden sich im Netzliteratur-Wiki unter https://wwik-prod.dla-marbach.de/line/index.php/Materialien_zur_Fortbildungsveranstaltung.

Links

Heritrix User Guide

(https://webarchive.jira.com/wiki/display/Heritrix/Heritrix+3.0+and+3.1+User+Guide)

pyWB (https://github.com/ikreymer/pywb)

ArchiveReady (http://archiveready.com/)

Wayback Machine (https://github.com/internetarchive/wayback)

Zusätzliche Literaturempfehlungen

Banos, Vangelis: *A quantitative approach to evaluate Website Archivability using the CLEAR+ method* in: *International Journal on Digital Libraries*, Berlin Heidelberg 2015.

Fritz, Steffen: *Praxisreport: Verfahren zur Evaluierung der Archivierbarkeit von Webobjekten* in ABI Technik 35 (2015), 117-120.

Masanès, Julien (Ed.): *Web Archiving*, Berlin 2006.

Stephanie Kuch
kuch@dla-marbach.de

Steffen Fritz
fritz@dla-marbach.de